

# 基于 BERT 的水稻表型知识图谱实体关系抽取研究

袁培森<sup>1</sup> 李润隆<sup>1</sup> 王 翀<sup>2</sup> 徐焕良<sup>1</sup>  
(1. 南京农业大学人工智能学院, 南京 210095; 2. 国网江苏省电力有限公司信息通信分公司, 南京 210024)

**摘要:** 针对水稻表型知识图谱中的实体关系抽取问题, 根据植物本体论提出了一种对水稻的基因、环境、表型等表型组学实体进行关系分类的方法。首先, 获取水稻表型组学数据, 并进行标注和分类; 随后, 提取关系数据集中的词向量、位置向量及句子向量, 基于双向转换编码表示模型 (BERT) 构建水稻表型组学关系抽取模型; 最后, 将 BERT 模型与卷积神经网络模型、分段卷积网络模型进行结果比较。结果表明, 在 3 种关系抽取模型中, BERT 模型表现更佳, 精度达 95.11%、F1 值为 95.85%。

**关键词:** 水稻表型; 知识图谱; 关系抽取; 双向转换编码表示模型  
中图分类号: TP391      文献标识码: A      文章编号: 1000-1298(2021)05-0151-08      OSID: 

## Entity Relationship Extraction from Rice Phenotype Knowledge Graph Based on BERT

YUAN Peisen<sup>1</sup> LI Runlong<sup>1</sup> WANG Chong<sup>2</sup> XU Huanliang<sup>1</sup>  
(1. College of Artificial Intelligence, Nanjing Agricultural University, Nanjing 210095, China  
2. Information and Communication Branch of State Grid Jiangsu Electric Power Co., Ltd., Nanjing 210024, China)

**Abstract:** Rice phenotype has an important guiding role in rice research by analyzing genetic information of various phenotype data. Knowledge graph technology has been widely used in knowledge storage and search engines by structurally describing the information, concepts, entities and relationships in data. As a key task in the knowledge graph, the relation extraction task can extract the connection between two entity words in the text. Within this research, rice phenotypic data was collected from the National Rice Data Center, and the data were preprocessed and annotated. The rice phenotype relationship was proposed based on the plant ontology, and then method of bidirectional encoder representation from transformers (BERT) was applied for classifying relation between rice genomics, environment, and phenotype data based on plant ontology. Then the word vector, position vector and sentence vector were extracted in the relation dataset, and rice phenotype relation extraction model was realized based on BERT. Finally, the results of BERT model was compared with the convolutional neural network and the piecewise convolutional network model. In the comparison of the three relationship extraction models, BERT achieved the best performance, and reached an accuracy of 95.11% and F1 value of 95.85%. Deep learning methods were used to improve the performance of relation extraction of knowledge graphs, which can provide technical support for the efficient construction of a rice phenotype knowledge graph system.

**Key words:** rice phenotype; knowledge graph; relationship extraction; bidirectional encoder representation from transformers

### 0 引言

植物表型组学数据分析是近年来植物学、信息科学领域研究的交叉热点,其本质是对植物基因数

据的三维时序表达,以及地域分布特征和代际演进规律<sup>[1]</sup>。表型组学指利用生物的遗传基因组信息对生物的外部及内部表型数据进行研究的一门综合性学科<sup>[2]</sup>。植物表型组学不仅研究植物的外在形

收稿日期: 2020-06-12    修回日期: 2020-12-02  
基金项目: 国家自然科学基金项目(61502236、61806097)和大学生创新创业训练专项计划项目(S20190025)  
作者简介: 袁培森(1980—),男,讲师,博士,主要从事智能信息处理和表型数据分析研究,E-mail: peiseny@njau.edu.cn  
通信作者: 徐焕良(1963—),男,教授,博士生导师,主要从事农业信息化与大数据技术研究,E-mail: huanliangxu@njau.edu.cn

状,还研究其内部结构、物理和生化性质以及遗传信息。亟需研究建立植物表型组学数据完整知识库的智能计算方法<sup>[3]</sup>。

中国是世界上水稻产量最大、消费最多的国家<sup>[4]</sup>,水稻的培育及研究是中国粮食安全战略的重要内容<sup>[5]</sup>。水稻表型组学研究是植物生物学的研究热点,水稻表型数据的高通量、高维、海量的数据特征对数据的快速检索和知识的有效提取提出了更高的技术要求<sup>[6]</sup>。

知识图谱将知识转化为图,利用计算机进行推理分析,实现从感知智能到认知智能的飞跃,是人工智能领域的一项重要技术<sup>[7]</sup>。知识图谱是一个具有结构化特征的语义知识库,采用符号的形式描述数据中的实体及之间的关系<sup>[8]</sup>,利用对语义的抽取和分析,并结合数据科学、人工智能等学科的前沿技术和方法,在学科知识库构建领域获得了广泛关注。

对知识图谱系统的构建包括 2 个核心步骤:实体抽取、实体间关系的构建,其中实体关系的构建需要关系的抽取技术。关系抽取任务的研究目标是自动对两个实体和之间联系所构成的 3 元组进行关系识别<sup>[9]</sup>。关系抽取能够提取文本数据中的特征,并提升到更高的层面<sup>[10]</sup>。

实体关系的抽取方法可以分为 3 类:基于模板、基于传统机器学习和基于深度学习的方法<sup>[11]</sup>。基于模板的关系抽取方法是早期基于语料学知识及语料的特点,由相应领域的专家和研究人員手工编写模板,这种方法需要耗费大量专业人力,可移植性较差。基于传统机器学习的关系抽取方法主要包括使用核函数<sup>[12]</sup>、逻辑回归<sup>[13]</sup>以及条件随机场<sup>[14]</sup>等,是一种依赖特征工程的方法。HASEGAWA 等<sup>[15]</sup>使用聚类方法计算上下文的相似性。赵明等<sup>[16]</sup>采用本体学习,使用有监督的、基于依存句法分析的词汇-语法模式对百度百科植物语料库进行关系抽取,在非分类的关系抽取任务中表现较好,为构建植物领域知识图谱奠定了基础。

基于深度学习的关系抽取方法包括递归神经网络模型<sup>[17]</sup>、卷积神经网络模型<sup>[18]</sup>、双向转换编码表示模型(Bidirectional encoder representation from transformers, BERT)<sup>[19]</sup>等。深度学习能够实现语义特征的自动提取,从而使模型能够对不同抽象层次上的语义进行分析<sup>[20]</sup>。BERT 为典型的深度学习模型<sup>[19]</sup>,通过自动学习句中特征信息、获取句子向量表示,能够对水稻表型组学数据进行关系抽取。在水稻知识图谱构建中,区分水稻表型组学实体之间的复杂关系与水稻表型组学知识库的构建有关。因此,研究水稻表型组学的关系抽取十分重要。

本文使用爬虫框架获取水稻表型组学数据,根据植物本体论提出一种对水稻的基因、环境、表型等表型组学数据进行关系分类的方法。使用词向量、位置向量等算法提取句中特征,在获取水稻表型组学实体关系数据集的基础上构建基于双向转换编码表示的关系抽取模型,并将本文方法与卷积神经网络(Convolutional neural network, CNN)<sup>[21]</sup>、分段卷积神经网络(Piece wise CNN, PCNN)<sup>[18]</sup>进行对比,以期实现句子级别的关系抽取。

## 1 水稻表型组学关系数据集

### 1.1 关系数据获取

本文关系数据集主要来自国家水稻数据中心(<http://www.ricedata.cn/>)以及维基百科中文语料库。数据爬取使用可对网页的结构性数据进行获取以及保存的框架 Scrapy<sup>[22]</sup>,实现水稻数据中心本体系统以及维基关系数据集的爬取。对爬取的水稻表型数据进行清洗处理,获得了用于关系分类处理的水稻表型组学关系数据集,数据集详情如表 1 所示。

表 1 数据集来源分布  
Tab. 1 Distribution of dataset origin

| 来源       | 数量    | 数据格式                           |
|----------|-------|--------------------------------|
| 水稻表型本体系统 | 2 812 | 术语名称 + 术语 ID + 术语定义 + 父术语 + 链接 |
| 维基百科语料库  | 3 135 | 实体 1 + 实体 2 + 关系 + 句子          |

### 1.2 关系数据分类

在水稻表型组学关系数据的分类问题上,本文参照了植物本体论(Plant ontology)<sup>[23]</sup>对植物表型组学的分类,通过关系分类将水稻的解剖结构、形态、生长发育与植物基因数据联系起来,从而对水稻表型组学数据进行分类。

本体<sup>[24]</sup>指的是在某一领域内的实体与其相互间关系的形式化表达,本体论是概念化的详细说明,它的核心作用是定义某一个领域内的专业词汇以及他们之间的关系<sup>[25]</sup>。

植物本体论<sup>[23]</sup>是一种结构化的数据库资源,是用来描述植物解剖学、形态学等植物学的结构性术语集合,它将植物的内部解剖结构、外表形态结构等表型组学数据与植物基因组学数据联系起来,使用关系来描述基因、环境、表型之间的联系。如今植物本体论的描述范围从最开始的水稻单个物种扩大到了 22 种植物,对这些植物的基因或基因模型、蛋白质、RNA、种质等表型和基因数据进行描述。本文依据其分类规则,将水稻表型组学数据分为 7 类:①is a,用来表示父术语以及子术语之间的关系,表

示对象 O1 是 O2 的子类型或亚型。②has part,用来表示对象 O1 的每个实例都有一部分 O2 的实例。③has a morphology trait,表示 O1 通过 O2 的形态特征表现出来。④develop from,表示 O1 从 O2 发育而来,O2 的世系可以追溯到 O1。⑤participate,表示实体 O1 的每个实例都参与开发 O2 的某些实例。⑥regulate,O1 对 O2 有调节或调控作用。⑦other,表示其他关系。

分类完成后的关系抽取数据集示例如表 2 所示。表 2 中,ddu1 (Dwarf and disproportionate uppermost-internode1)为使用甲基磺酸乙酯诱变粳稻品种兰胜而成的矮化突变体的品种名称;SPL5 (Spotted leaf 5)为经  $\gamma$  射线辐射诱变粳稻品种 Norin 8 而成的水稻类病变突变体的品种名称;FLW1 (Flag leaf width NAL1)为剑叶宽度基因。最后,将数据集按 8:2 分为训练集和测试集。

表 2 关系抽取数据集示例

| Tab.2 Examples of relation extraction datasets |         |                        |                                 |
|------------------------------------------------|---------|------------------------|---------------------------------|
| 实体 1                                           | 实体 2    | 关系                     | 句子                              |
| 产量性状                                           | 性状      | is a                   | 产量性状表示与植物的可收获产物相关的性状            |
| 内保护层                                           | 果皮      | has part               | 大米具有内保护层,包括了果皮、种皮以及珠心层          |
| ddu1                                           | 矮化      | has a morphology trait | 突变体植株 ddu1 表现出明显的矮化             |
| 胚芽                                             | 花药      | develop from           | 花药与子房中间的胚珠相互结合,发育形成胚芽           |
| SPL5                                           | RNA 的剪接 | participate            | SPL5 参与植物 RNA 的剪接,负调控细胞死亡以及抗性应答 |
| FLW1                                           | 剑叶宽度    | regulate               | 剑叶宽度受到两个显性基因的控制,即 FLW1、FLW2     |
| 稻穗                                             | 穗伸出度    | other                  | 穗伸出度是用来衡量稻穗从旗叶叶鞘中伸出长度的指标        |

1.3 关系数据存储

水稻实体及关系采用图方式进行建模以及数据存储,本文使用图数据库 Neo4j<sup>[26]</sup> 存放实体和关系数据。Neo4j 的核心概念是节点和边,节点用来存储实体,使用圆形图例表示,边用来存储关系数据结构中实体之间的关系,使用带箭头的线表示。不同实体以及关系的相互连接形成复杂的数据结构,实现对某个实体进行关系的完整增删改查等功能。

对收集的数据集进行预处理,提取 2 021 个实体和 2 689 条关系,通过 Cypher 语言<sup>[27]</sup> 进行快速的查询工作。图 1 为 Neo4j 数据库存储的水稻表型组学关系示例。由于实体名称较长,图 1 中的“12 号染色...”为 12 号染色体;“等位基因...”为等位基因

STV11 - S。

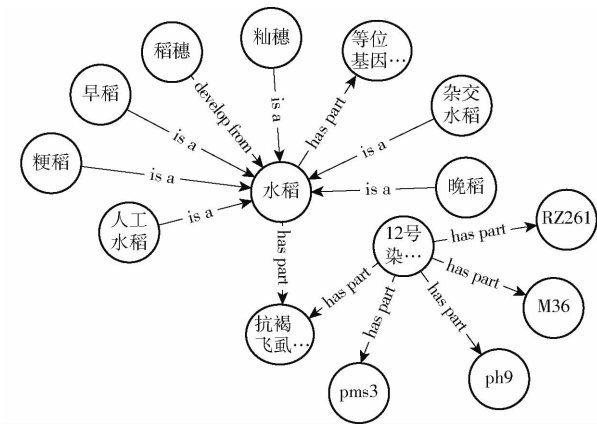


图 1 Neo4j 水稻表型关系示例

Fig.1 Example in Neo4j of rice phenotypic relationship

2 水稻表型组学关系抽取模型构建

2.1 向量化表示

本文 BERT 关系抽取模型使用词向量、位置向量以及句子向量相结合的输入向量序列,不仅能简单获取词语语义上的特征,而且能够对深层次语义进行表示和抽取。

2.1.1 词向量

本文使用 BERT 模型中的词嵌入方式来动态产生词向量,即将词转化为稠密的向量。通过这种词嵌入方式,该模型能够根据上下文预测中心词的方式来获得动态的语义特征,以解决传统词嵌入模型产生的多义词局限性,可以产生更精确的特征表示,从而提高模型性能。

BERT 的词向量生成方法如下:给定语句序列  $s = w_0, w_1, \dots, w_n$ 。其中  $w_0 = [\text{CLS}]$ 、 $w_n = [\text{SEP}]$  表示句子的开始以及结束。模型将原有的序列映射为具有固定长度的向量来表示语义关系。

2.1.2 位置向量

设句子为  $s = w_0, w_1, \dots, w_n$ , 实体为  $i_1$  与  $i_2$ , 则对于每一个单词  $w_i$ , 计算其与  $i_1$ 、 $i_2$  的相对距离,即  $i - i_1$  和  $i - i_2$ , 使得该句子可以根据两个实体生成两部分的位置向量,并且能体现距离和实体的关系。本文使用的位置向量维度为 50。

2.1.3 句子向量

句子向量按照句子的数目进行标记,对于第 1 条句子的每个单词添加向量  $v_1$ , 给第 2 条句子中的每个单词添加一个向量  $v_2$ 。

2.1.4 输入表示

BERT 模型的输入示例如图 2 所示。图 2 中的 BERT 模型输入的句子为“稻是谷类,原产中国与印度”,模型生成每个词的词向量,根据每个词与实

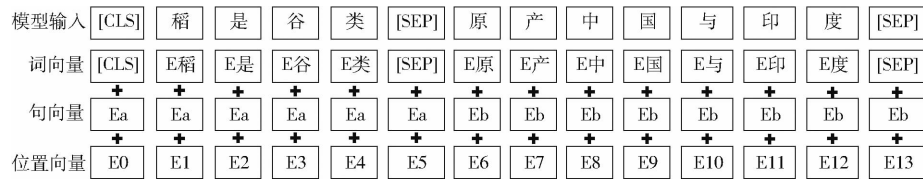


图 2 BERT 模型的输入示例

Fig. 2 Example of input of BERT model

体之间的距离生成句向量,根据句子的条数生成对应的句向量,将此作为 BERT 模型的输入。

2.2 BERT 关系抽取模型构建

BERT 是以 Transformer 的编码器为基础的双向自注意力机制表示模型,能够对所有层基于上下文进行双向表示。BERT 模型使用双向自注意力机制来进行构建,使用 Transformer 的编码器来进行编码,并且使用遮挡语言模型以及下一句预测两个方法来更有效地训练模型。

2.2.1 双向自注意力机制

BERT 使用双向自注意力机制<sup>[28]</sup>进行构建。双向自注意力机制是注意力机制中的一种,注意力机制在自然语言处理领域的多个任务得到了实际应用。注意力机制可以描述为一个查询  $Q$  到相应键值对  $\langle K, V \rangle$  的一个映射过程<sup>[29]</sup>,可描述为

$$A_i(Q, K, V) = S_f(S_m(Q, K))V \tag{1}$$

式中  $A_i$ ——注意力机制函数  
 $S_f$ ——Softmax 函数  
 $S_m$ ——相似度函数

注意力值的计算过程可分为 3 部分:①首先计算查询  $Q$  和每个键  $K$  之间的相似度  $S$ ,获得权重,使用的相似度计算函数有点积、拼接以及感知机。②使用 Softmax 函数进行权重归一化。③将权重以及键值对中的值  $V$  进行加权,获得最终的注意力值。自注意力机制即检索自身的键值对进行加权处理, $Q = K = V$ ,将序列进行重新编码,获得更具整体性的特征序列<sup>[30]</sup>。自注意力机制的结构图如图 3 所示。

自注意力机制将输入序列通过向量映射的方式输入到嵌入层,注意力层进行查询向量和值向量的相似度计算,Softmax 层使用函数加权后将序列输出。BERT 所用的多头自注意力机制在输入到注意力层之前对查询  $Q$ 、键  $K$  以及值  $V$  进行多次线性变换,线性变换的次数即为多头,多头自注意力机制可以获得多种序列的子特征,进而获得较长序列中的相隔较远的向量特征<sup>[31]</sup>。

2.2.2 Transformer 编码器

BERT 使用 Transformer 编码器进行编码,Transformer<sup>[32]</sup>通过对语义信息以及位置信息的分析

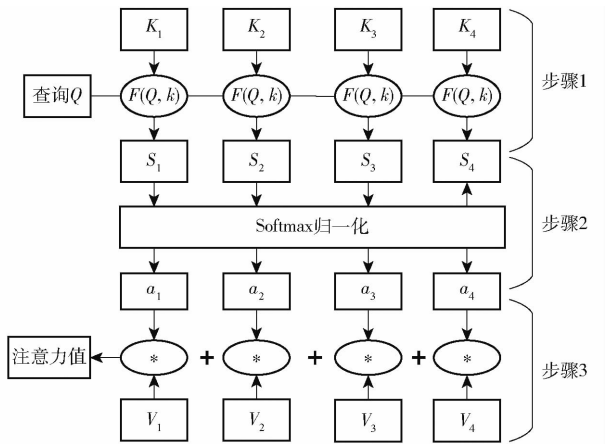


图 3 自注意力机制结构图

Fig. 3 Self-attention mechanism structure diagram

来完成自然语言处理任务,其框架为编码器加解码器结构。其中,编码器框架使用了层叠结构,每一层有两部分:进行加权处理的多头注意力机制和进行前馈化网络的全连接层,在两部分之间使用残差进行连接然后进行标准化。解码器的层数与编码器相同,同时在每一层之内还添加了一个进行计算翻译效果的部分。Transformer 编码器结构图如图 4 所示,图中  $N \times$  表示编码器或解码器包含的层数。

Transformer 编码器由 3 部分组成:①首先对输入句子进行向量化,将词嵌入到编码器中。②编码器接受向量序列,随后使用自注意力机制对序列进行处理,通过对序列中所有单词之间建立联系来进行序列编码,处理后的序列通过残差网络进行求和与归一化。③自注意力机制结束以后,输入到全连接的前馈网络中,输出标准化后的向量。

BERT 模型使用多个 Transformer 编码器进行编码,编码器输出后进入到一个全连接层与激活函数构成的分类层并输出相应的概率<sup>[33]</sup>。图 5 是对水稻表型进行编码示例,输入的句子为“产量性状是与植物可收获产物相关的性状”。

BERT 模型在使用过程中,仅需要在编码器后面加上一层全连接层就能够完成关系抽取任务。在后期的微调部分中,设之前遮盖处理后的输出向量为  $C$ ,使用 Softmax 分类器完成关系分类的概率  $P_r$  为

$$P_r = S_f(CW^T) \tag{2}$$

式中  $W$ ——向量矩阵

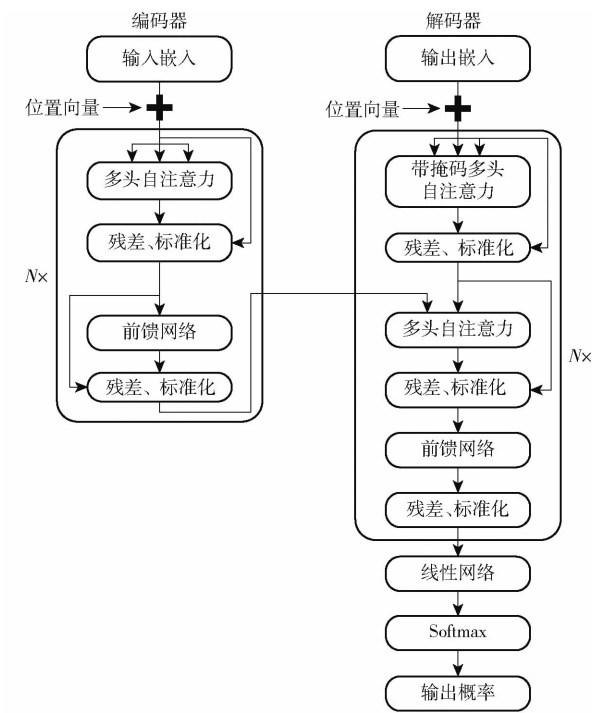


图 4 Transformer 结构图

Fig. 4 Transformer structure diagram

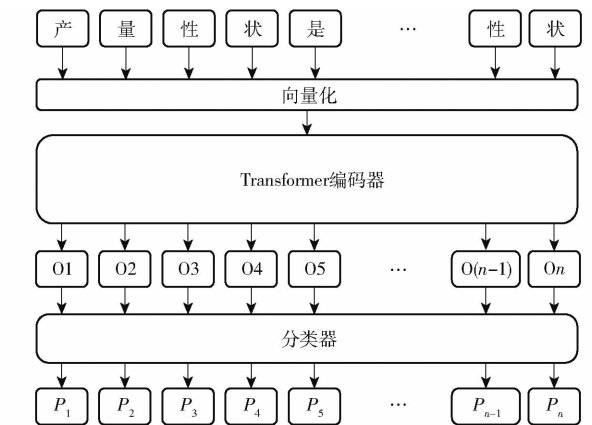


图 5 BERT 模型结构图

Fig. 5 BERT model structure diagram

对于本文的关系多分类问题,类别标签  $y \in \{1, 2, \cdots, M\}$ 。给定测试样本  $x$ , Softmax 函数预测类别  $c \in \{1, 2, \cdots, M\}$  的条件概率为

$$p(y = c | x) = S_f(\mathbf{w}_c^T x) = \frac{\exp(\mathbf{w}_c^T x)}{\sum_{i=1}^M \exp(\mathbf{w}_i^T x)} \quad (3)$$

式中  $\mathbf{w}_c$ ——权重  
 $\mathbf{w}_i$ ——第  $i$  类权重  
 $p$ ——概率

BERT 模型输出关系类别以及其对应的概率。另外, BERT 模型在预训练部分使用了遮挡语言模型以及下一句预测两个方式来训练模型。

2.2.3 遮挡语言模型

遮挡语言模型 (Masked language model)<sup>[19]</sup>指的是在进行 BERT 模型训练时,由于进行的注意力机

制是多头而不是单向的,如果按照 CNN 等模型的训练方式进行训练,则 BERT 模型的训练将成为一个先获得后文再进行预测的任务,无法正确获取语义特征,因此进行双向注意力机制训练时, BERT 使用了遮挡语言模型,将输入的词进行随机遮盖,从而使得双向编码器能够真正对前后文进行预测<sup>[19]</sup>。本文对 15% 的词进行遮挡,并且遵循以下规律:①被遮挡的词有 80% 的概率被替换成屏蔽符号 [mask]。②10% 的概率被换成随机词。③10% 的概率保持原有单词不变。这样后期微调部分的向量输入不会与遮盖处理中的向量差距太大。

2.2.4 下一句预测

下一句预测 (Next sentence prediction)<sup>[19]</sup>使 BERT 模型能够学习下一句和上一句的内在联系, BERT 模型在数据集中随机选取句子  $S_1$ , 对于其下一句  $S_2$ , 有 50% 的概率将  $S_2$  替换为无关的句子  $S_3$ , 以此来学习句子间的关系。

3 关系分类结果与分析

3.1 试验环境

选择 Intel Corei5 - 8250u 处理器 @ 1.6 GHz, 8 GB 内存, 1 TB 硬盘, Windows 10 操作系统。

3.2 参数设置

BERT 模型的参数设置如表 3 所示。为防止模型训练后期的波动,学习率衰减采用了文献[34]中的推荐值,设置为  $2 \times 10^{-5}$ 。

表 3 BERT 模型参数设置  
Tab.3 BERT parameter settings

| 参数     | 数值/算法              |
|--------|--------------------|
| 隐藏层数量  | 1 536              |
| 学习率衰减  | $2 \times 10^{-5}$ |
| 最大序列长度 | 80                 |
| 迭代次数   | 5                  |
| 批尺寸    | 8                  |
| 梯度下降算法 | ADAMW              |

梯度下降算法 (Gradient descent optimizer)<sup>[35]</sup>能够帮助模型进行目标函数的最大化或最小化计算,一个优秀的梯度下降算法能够减少损失函数的值。常用的梯度下降算法有随机梯度下降 (Stochastic gradient descent, SGD)<sup>[35]</sup>、自适应力矩估计 (Adaptive moment estimation, ADAM)<sup>[36]</sup>、解耦权重衰减的自适应矩估计 (Adaptive moment estimation with decoupled weight decay, ADAMW)<sup>[37]</sup>等,本文选择 ADAMW 算法。

3.3 数据集

根据植物本体论进行实体关系数据的分类,共

获得 7 大类、2 689 条关系数据,类型有:is a、has part、has a morphology trait、develop from、participate、regulate、other。各个关系类型的数量及分布如表 4 所示。

表 4 水稻表型组学关系数据集的数量分布  
Tab.4 Distribution of relationship quantity of rice phenotypic relationship in dataset

| 关系类型                   | 数量    |
|------------------------|-------|
| is a                   | 1 088 |
| has part               | 456   |
| has a morphology trait | 152   |
| develop from           | 127   |
| participate            | 160   |
| regulate               | 235   |
| other                  | 471   |

3.4 算法性能评估指标

使用精度 (Precision,  $P$ )、召回率 (Recall,  $R$ )、F1 值 ( $F_1$ ) 作为评价指标,将 BERT 与传统的卷积神经网络模型<sup>[21]</sup>与分段卷积神经网络模型<sup>[18]</sup>进行对比。

3.5 BERT 模型关系分类结果

本部分对梯度下降算法<sup>[35]</sup>、批尺寸<sup>[38]</sup>和表 2 中的关系进行了试验分析测试。

3.5.1 梯度下降算法

对于 BERT 关系抽取模型,本文进行了梯度下降算法的对比,选择批 (Batch) 尺寸为 8,3 种梯度下降算法在 BERT 模型上的结果如图 6 所示。

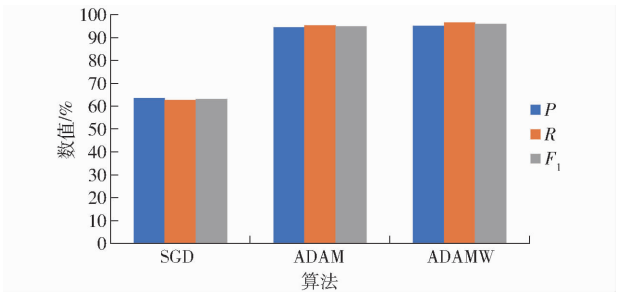


图 6 不同梯度下降算法在 BERT 模型上的结果  
Fig.6 Comparison of gradient descent algorithm on BERT model

由图 6 可以看出,ADAMW 的精度、召回率和 F1 值比 SGD 和 ADAM 高,SGD 最低,3 个指标均在 60% 左右。ADAM 和 ADAMW 都在 94% 以上。

3.5.2 批尺寸

选择批尺寸分别为 8、16、32、64 进行试验,选择 ADAMW 作为梯度下降算法,其在 BERT 模型上的结果如表 5 所示。

由表 5 可知,批尺寸为 8 时,ADAMW 算法的精度达到了 95.11%,召回率为 96.61%,F1 值为

表 5 不同批尺寸在 BERT 模型上的对比  
Tab.5 Comparison of different batch on BERT model

| 批尺寸 | $P$   | $R$   | $F_1$ |
|-----|-------|-------|-------|
| 8   | 95.11 | 96.61 | 95.85 |
| 16  | 94.59 | 95.02 | 94.81 |
| 32  | 94.48 | 96.61 | 95.63 |
| 64  | 94.23 | 95.02 | 94.62 |

95.85%。相比批尺寸为 16、32、64,精度分别提高了 0.52、0.63、0.88 个百分点;F1 值分别提高 1.04、0.22、1.23 个百分点。

3.5.3 不同关系类型的处理结果

本试验批尺寸为 8,BERT 模型使用 ADAMW 算法对本文数据集上的不同关系抽取结果进行对比,结果如表 6 所示。

表 6 BERT 模型对不同关系的处理结果  
Tab.6 Results of different relations in BERT model

| 关系类型                   | $P$   | $R$   | $F_1$ |
|------------------------|-------|-------|-------|
| develop from           | 96.07 | 43.64 | 60.02 |
| has a morphology trait | 96.67 | 48.33 | 64.44 |
| has part               | 86.14 | 82.86 | 84.47 |
| is a                   | 98.57 | 87.71 | 92.83 |
| participate            | 96.67 | 48.33 | 64.45 |
| regulate               | 91.67 | 61.11 | 73.33 |
| other                  | 91.31 | 77.78 | 83.99 |

由表 6 可知,BERT 模型对于不同关系的 F1 值都不低于 60.02%,但是对于不同关系的处理效果也不同。其中,对于 has part、is a、other、regulate 关系分类效果较好,其 F1 值都不小于 73.33%,而对于 develop from、participate、has a morphology trait 的分类效果相对较差。在 7 种关系中,is a 关系类型的测试结果最佳,其 F1 值达到了 92.83%,是 develop from 类型的 1.546 7 倍。develop from、has a morphology trait 和 participate 分类效果较差的原因是这 3 个关系类别的数据库中关系数较少,且数据集中各个类别的分布不均衡。其解决方法有:①通过增加这 3 个类别实体关系数据使 BERT 模型提取更多有效的语义和词汇特征。②将各个关系数据的条数进行调整,保持各个类别实体关系数据的数量均衡。

3.6 模型对比

本文将 CNN<sup>[21]</sup>、PCNN<sup>[18]</sup>与 BERT 模型进行对比,CNN 和 PCNN 模型的参数设置如表 7 所示。

CNN 在批尺寸为 16 时,使用 SGD 算法时获得最高精度、召回率与 F1 值,精度为 81.79%,召回率

表 7 CNN 和 PCNN 模型参数设置

Tab.7 CNN and PCNN model parameter settings

| 参数     | 数值/方法                |
|--------|----------------------|
| 卷积核尺寸  | 3                    |
| 隐藏层数   | 230                  |
| 填充值    | 1                    |
| 丢弃率    | 0                    |
| 词向量维度  | 100                  |
| 权重衰减   | $1 \times 10^{-5}$   |
| 学习率衰减  | 0.1                  |
| 卷积激活函数 | ReLU <sup>[39]</sup> |
| 池化方法   | 最大池化                 |
| 迭代次数   | 5                    |

为 82.35% ,F1 值为 82.07% 。PCNN 的批尺寸为 16,使用 SGD 算法时,获得最高精度、召回率与 F1 值,精度为 85.95% ,召回率为 81.67% ,F1 值为 83.66% 。BERT 模型在隐藏层数量为 1 536、最大序列长度为 80、学习率衰减为  $2 \times 10^{-5}$ 、训练轮数为 5、批尺寸为 8、梯度下降算法为 ADAMW 时,关系抽取的精度、召回率与 F1 值达到最优,精度为 95.11% ,

召回率为 96.61% ,F1 值为 95.85% 。

BERT 在精度、召回率以及 F1 值上都明显高于其他两种模型,其 F1 值是 CNN 的 1.17 倍、PCNN 的 1.15 倍。

综上所述,在使用 BERT 模型进行水稻表型组学数据关系抽取时,BERT 模型能够根据上下文预测中心词的方式来获得动态的词向量,使用自注意力机制获得双向的语义特征,大幅度提高了关系抽取的质量。

4 结束语

本文基于植物本体论提出基于水稻表型组学的关系分类方法,将水稻表型的实体关系分为 7 类,使用词向量、位置向量以及句子向量进行句子特征抽取,构建 BERT 模型,并将 BERT 模型与 CNN、PCNN 模型进行对比。结果表明,BERT 模型的精度、召回率与 F1 值分别为 95.11%、96.61% 和 95.85% ,达到了预期分类效果。

参 考 文 献

[1] 张慧春,周宏平,郑加强,等. 植物表型平台与图像分析技术研究进展与展望[J/OL]. 农业机械学报,2020,51(3):1-17. ZHANG Huichun, ZHOU Hongping, ZHENG Jiaqiang, et al. Research progress and prospect in plant phenotyping platform and image analysis technology[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2020,51(3):1-17. [http://www.j-csam.org/jcsam/ch/reader/view\\_abstract.aspx?flag=1&file\\_no=20200301&journal\\_id=jcsam](http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20200301&journal_id=jcsam). DOI:10.6041/j.issn.1000-1298.2020.03.001. (in Chinese)

[2] 周济,TARDIEU F,PRIDMORE T,等. 植物表型组学:发展、现状与挑战[J]. 南京农业大学学报,2018,41(4):580-588. ZHOU Ji, TARDIEU F,PRIDMORE T, et al. Plant phenomics: history, present status and challenges[J]. Journal of Nanjing Agricultural University, 2018, 41(4): 580-588. (in Chinese)

[3] 潘映红. 论植物表型组和植物表型组学的概念与范畴[J]. 作物学报,2015,41(2):175-186. PAN Yinghong. Analysis of concepts and categories of plant phenome and phenomics [J]. Acta Agronomica Sinica, 2015, 41(2): 175-186. (in Chinese)

[4] 凌霄霞,张作林,翟景秋,等. 气候变化对中国水稻生产的影响研究进展[J]. 作物学报,2019,45(3):323-334. LING Xiaoxia, ZHANG Zuolin, ZHAI Jingqiu, et al. A review for impacts of climate change on rice production in China[J]. Acta Agronomica Sinica,2019,45(3):323-334. (in Chinese)

[5] 黎志康.我国水稻分子育种计划的策略[J]. 分子植物育种,2005(5):603-608. LI Zhikang. Strategies for molecular rice breeding in China[J]. Molecular Plant Breeding, 2005(5): 603-608. (in Chinese)

[6] 陈凯文,俞双恩,李倩倩,等. 不同水文年型下水稻节水灌溉技术方案模拟与评价[J/OL]. 农业机械学报,2019,50(12):268-277. CHEN Kaiwen,YU Shuang'en, LI Qianqian, et al. Simulation and evaluation of technical schemes for water-saving irrigation of rice in different hydrological years[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2019,50(12):268-277. [http://www.j-csam.org/jcsam/ch/reader/view\\_abstract.aspx?flag=1&file\\_no=20191231&journal\\_id=jcsam](http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20191231&journal_id=jcsam). DOI:10.6041/j.issn.1000-1298.2019.12.031. (in Chinese)

[7] 漆桂林,高桓,吴天星. 知识图谱研究进展[J]. 情报工程,2017,3(1):4-25. QI Guilin, GAO Huan, WU Tianxing. The research advances of knowledge graph [J]. Technology Intelligence Engineering, 2017, 3(1): 4-25. (in Chinese)

[8] 刘知远,孙茂松,林衍凯,等. 知识表示学习研究进展[J]. 计算机研究与发展,2016,53(2):247-261. LIU Zhiyuan, SUN Maosong, LIN Yankai, et al. Knowledge representation learning: a review [J]. Journal of Computer Research and Development, 2016, 53(2): 247-261. (in Chinese)

[9] SCHUTZ A, BUITELAAR P. RelExt : a tool for relation extraction from text in ontology extension[C]//International Semantic Web Conference, Galway, Ireland,2005:593-606.

[10] 郭喜跃,何婷婷. 信息抽取研究综述[J]. 计算机科学,2015,42(2):14-17. GUO Xiyue, HE Tingting. Survey about research on information extraction[J]. Computer Science, 2015, 42(2): 14-17. (in Chinese)

[11] 鄂海红,张文静,肖思琪,等. 深度学习实体关系抽取研究综述[J]. 软件学报,2019,30(6):1793-1818. E Haihong, ZHANG Wenjing, XIAO Siqi, et al. Survey of entity relationship extraction based on deep learning [J]. Journal of Software,2019,30(6):1793-1818. (in Chinese)

[12] LIN Y, LIU Z, SUN M. Neural relation extraction with multi-lingual attention[C]//Proceedings of the 55th Annual Meeting of

the Association for Computational Linguistics, Vancouver, Canada, 2017:34 – 43.

- [13] WANG M. A re-examination of dependency path kernels for relation extraction[C] // Proceedings of the Third International Joint Conference on Natural Language Processing, Hyderabad, India, 2008:841 – 846.
- [14] CULOTTA L, GIANGUZZA A, MANNINO M R, et al. Polycyclic aromatic hydrocarbons (pah) in Vulcano Island (Aeolian Archipelago) mud utilized for therapeutic purpose[J]. Polycyclic Aromatic Compounds, 2007, 27(4): 281 – 294.
- [15] HASEGAWA T, SEKINE S, GRISHMAN R. Discovering relations among named entities from large corpora[C] // Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics: Association for Computational Linguistics, Sapporo, Japan, 2004: 415 – 423.
- [16] 赵明,杜亚茹,杜会芳,等. 植物领域知识图谱构建中本体非分类关系提取方法[J/OL]. 农业机械学报, 2016, 47(9): 278 – 284.  
ZHAO Ming, DU Yaru, DU Huifang, et al. Research on ontology non-taxonomic relations extraction in plant domain knowledge graph construction [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(9):278 – 284. [http://www.j-csam.org/jcsam/ch/reader/view\\_abstract.aspx?flag=1&file\\_no=20160938&journal\\_id=jcsam](http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20160938&journal_id=jcsam). DOI: 10.6041/j.issn.1000-1298.2016.09.038. (in Chinese)
- [17] KATIYAR A, CARDIE C. Going out on a limb: joint extraction of entity mentions and relations without dependency trees[C] // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Honolulu, Hawaii, USA, 2017: 917 – 928.
- [18] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[C] // Proceedings of the 25th International Conference on Computational Linguistics, Dublin, Ireland, 2014: 2335 – 2344.
- [19] DEVLIN J, CHANG M, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv, 2018:1810.04805.
- [20] 袁培森,杨承林,宋玉红,等. 基于 Stacking 集成学习的水稻表型组学实体分类研究[J/OL]. 农业机械学报, 2019, 50(11): 144 – 152.  
YUAN Peisen, YANG Chenglin, SONG Yuhong, et al. Classification of rice phenomics entities based on Stacking ensemble learning[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(11):144 – 152. [http://www.j-csam.org/jcsam/ch/reader/view\\_abstract.aspx?flag=1&file\\_no=20191116&journal\\_id=jcsam](http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20191116&journal_id=jcsam). DOI:10.6041/j.issn.1000-1298.2019.11.016. (in Chinese)
- [21] HAFEMANN L G, SABOURIN R, OLIVEIRA L S. Learning features for offline handwritten signature verification using deep convolutional neural networks[J]. Pattern Recognition, 2017, 70: 163 – 176.
- [22] WANG J, GUO Y. Scrapy-based crawling and user-behavior characteristics analysis on taobao[C] // 2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Sanya, China, 2012: 44 – 52.
- [23] COOPER L, JAISWAL P. The plant ontology: a tool for plant genomics[M]. Plant Bioinformatics, Springer, 2016: 89 – 114.
- [24] 唐晓波,翟夏普. 基于本体知识集合的知识检索研究[J]. 图书馆学研究, 2018(1): 60 – 66.  
TANG Xiaobo, ZHAI Xiapu. A research on the knowledge retrieval based on ontology knowledge set [J]. Research on Library Science, 2018(1): 60 – 66. (in Chinese)
- [25] MAEDCHE A, STAAB S. Ontology learning for the semantic web[J]. IEEE Intelligent Systems, 2001, 16(2): 72 – 79.
- [26] WEBBER J. A programmatic introduction to Neo4j[C] // Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity, Chicago, IL, USA, 2012: 217 – 218.
- [27] HOLZSCHUHER F, PEINL R. Performance of graph query languages: comparison of cypher, gremlin and native access in Neo4j[C] // Proceedings of the Joint EDBT/ICDT 2013 Workshops, Genoa, Italy, 2013: 195 – 204.
- [28] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[J]. arXiv preprint arXiv, 2016:1611.01603.
- [29] XIAO T, XU Y, YANG K, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, Massachusetts, USA, 2015: 842 – 850.
- [30] 胡荣磊,芮璐,齐筱,等. 基于循环神经网络和注意力模型的文本情感分析[J]. 计算机应用研究, 2019, 36(11): 3282 – 3285.  
HU Ronglei, RUI Lu, QI Xiao, et al. Text sentiment analysis based on recurrent neural networks and attention model[J]. Application Research of Computers, 2019, 36(11): 3282 – 3285. (in Chinese)
- [31] 袁和金,张旭,牛为华,等. 融合注意力机制的多通道卷积与双向 GRU 模型的文本情感分析研究[J]. 中文信息学报, 2019, 33(10): 109 – 118.  
YUAN Hejin, ZHANG Xu, NIU Weihua, et al. Sentiment analysis based on multi-channel convolution and bi-directional GRU with attention mechanism[J]. Journal of Chinese Information Processing, 2019, 33(10): 109 – 118. (in Chinese)
- [32] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C] // Advances in Neural Information Processing Systems, Long Beach, CA, USA, 2017: 5998 – 6008.
- [33] 沈忱林,张璐,吴良庆,等. 基于双向注意力机制的问答情感分类方法[J]. 计算机科学, 2019, 46(7): 151 – 156.  
SHEN Chenlin, ZHANG Lu, WU Liangqing, et al. Sentiment classification towards question-answering based on bidirectional attention mechanism. [J]. Computer Science, 2019, 46(7): 151 – 156. (in Chinese)
- [34] SMITH L N. A disciplined approach to neural network hyper-parameters: Part 1—learning rate, batch size, momentum, and weight decay[J]. arXiv preprint arXiv, 2018:89 – 132.
- [35] BOTTOU L. Large-scale machine learning with stochastic gradient descent[C] // Proceedings of COMPSTAT'2010, Springer, 2010: 177 – 186.
- [36] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. arXiv preprint arXiv, 2014:1 – 39.
- [37] LOSHCILLOV I, HUTTER F. Fixing weight decay regularization in adam[J]. arXiv preprint arXiv, 2018:14 – 28.
- [38] YOU Y, GITMAN I, GINSBURG B. Scaling SGD batch size to 32k for imagenet training[J]. arXiv preprint arXiv, 2017: 6 – 21.
- [39] HARA K, SAITO D, SHOUNO H. Analysis of function of rectified linear unit used in deep learning[C] // Proceedings of 2015 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2015:1 – 8.